Systematics and Biodiversity **1** (2): 139–143
DOI: 10.1017/S1477200003001178 Printed in the United Kingdom

Issued 14 August 2003

The *Big Machine* and the much-maligned taxonomist

Robert Scotland*
Colin Hughes
Donovan Bailey
& Alexandra Wortley
Department of Plant Sciences,
University of Oxford, South
Parks Road, Oxford,
OX1 3RB, UK

submitted January 2003
accepted April 2003

DNA taxonomy and the web

Every so often taxonomy enters the mainstream of biological discussion (Bisby *et al.*, 2002; Gewin, 2002; Godfray, 2002; Lee, 2002; Lipscomb *et al.*, 2003; Mallet & Willmott, 2003; Seberg *et al.*, 2003; Tautz *et al.*, 2002, 2003). Most recently, in response to a perceived crisis in taxonomy, two proposals to modify taxonomic practice have been put forward. First, Godfray (2002) has proposed that all new taxonomic revisions are placed on the web, available and accessible to all. In addition, he suggests drawing a line under nomenclatural issues, freeing taxonomists to concentrate on more substantial matters. Second, Tautz *et al.* (2002, 2003), as others before and since (Hebert *et al.*, 2003), have proposed that DNA sequences should be the central scaffold for taxonomy. Both DNA and web-based taxonomy have been touted as providing practical and technological solutions to a range of issues concerned with the global inventory of taxa.

Figure 1 illustrates the research programme of taxonomy, which seeks to document encyclopedic knowledge of all taxa on earth (Wilson, 2003). Recent commentary has suggested that this research programme is in crisis for four reasons. First, the huge shortfall in our current knowledge of taxonomic diversity. It is generally estimated that only around 10% of the world's biota has so far been described (Wilson, 2000). Second, the pace of progress towards completing the encyclopedia is too slow. For example, the Species Plantarum Project dating back to 1995 which aims to complete a taxonomic account of all vascular plants has in seven years completed only six families comprising 1100 species, a mere 0.25%. At this rate, this valiant and ambitious project will reach completion in 2800 years. What this reflects is the paucity of new taxonomic treatments covering whole families. Other measures of the rate of taxonomic progress across all groups of organisms (e.g. numbers of new species being described per year, or the number of monographs being written) show a similar glacial rate of advancement towards completing the inventory. Third, taxonomists spend a significant part of their time sorting out the often complex synonymy and scattered type material associated with 250 years of previous work (Godfray, 2002) rather than delimiting and describing taxa. Finally, and perhaps more cogently, the reason the encyclopedia of life is viewed as being in crisis is the destruction of habitats and extinction of species. This means that species may be lost before they can be described and that conservation priorities are constrained by what we know about a small subset of total biodiversity.

Recent discussions about the perceived crisis in taxonomy contrast traditional and technological approaches (Bisby *et al.*, 2002; Gewin, 2002; Godfray, 2002; Lee, 2002; Lipscomb *et al.*, 2003; Mallet & Willmott, 2003; Seberg *et al.*, 2003; Tautz *et al.*, 2002, 2003). Figure 2 is a caricature of the aforementioned technological solutions of DNA and web-based taxonomy and contrasts with the more rounded vision of taxonomic practice given in Fig. 1. In our opinion, the contrast is unnecessary and unhelpful. To place the elements of Fig. 2 at the centre of taxonomic practice is to over-inflate the significance of these important technological developments, all of which are readily assimilated into Fig. 1.

While most would accept that web-based repositories of taxonomic outputs are not a new idea and are already under construction (Gewin, 2002; Bisby *et al.*, 2002; Lee, 2002) and welcome, they provide no practical solution or significant answers to the shortfall in our knowledge of the global inventory of organisms. Undoubtedly, integration of database technology and the increasing availability of online databases of names, type specimens and taxonomic literature have increased the efficiency and productivity of taxonomy, and more gains are to be expected as these resources are further enhanced. However, these gains remain small in relation to the scale of the problem, and do not address the central tasks of taxonomy – sampling, data gathering and analysis, delimitation of taxa and classification (Fig. 1). Similarly misdirected is the idea that web-based taxonomic revisions can in some way draw a line under all earlier taxonomic literature such that it would no longer be relevant or necessary to consult it (Godfray, 2002). Godfray claims that such revisions, once completed, could serve as the unitary taxonomic focus for a particular group. At one level this is nothing new – because substantial revisions and monographs do already act as a unitary focus for the taxonomy of a group. The crucial point is that such revisions and monographs are few and far between. Godfray's suggestion amounts to not much more than making monographs available online. We believe that the greatest benefit of this will not be to speed up the completion of the global inventory, but rather to make it more democratic and the outputs more accessible, benefits that we wholeheartedly welcome and endorse.

We characterize DNA taxonomy as part of the *Big Machine* with high throughput DNA sequencing at its heart and

* Corresponding author.
Email: robert.scotland@plant-sciences.oxford.ac.uk

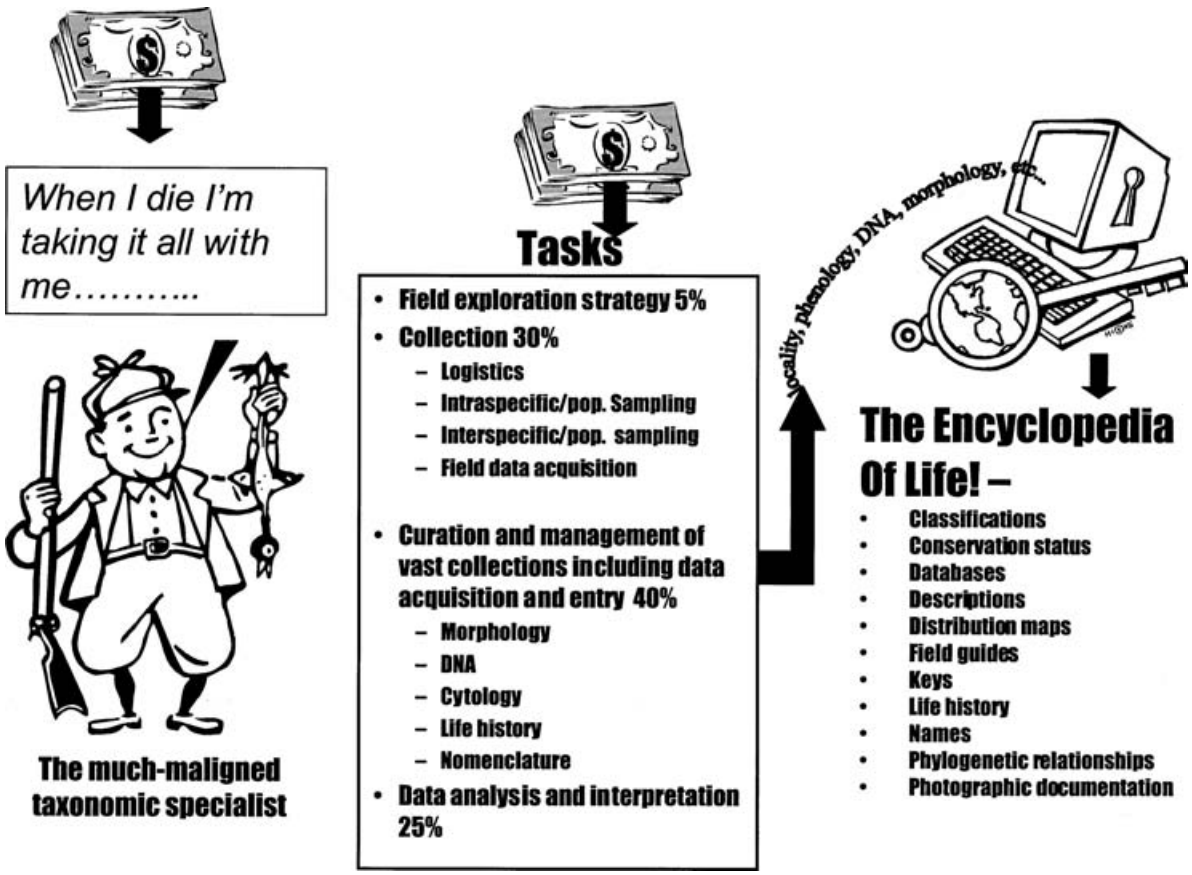

Figure 1 Cartoon illustrating the research programme of taxonomy, which seeks to document encyclopedic knowledge of all taxa on earth. Images from www.discoveryschool.com and www.barrysclipart.com.

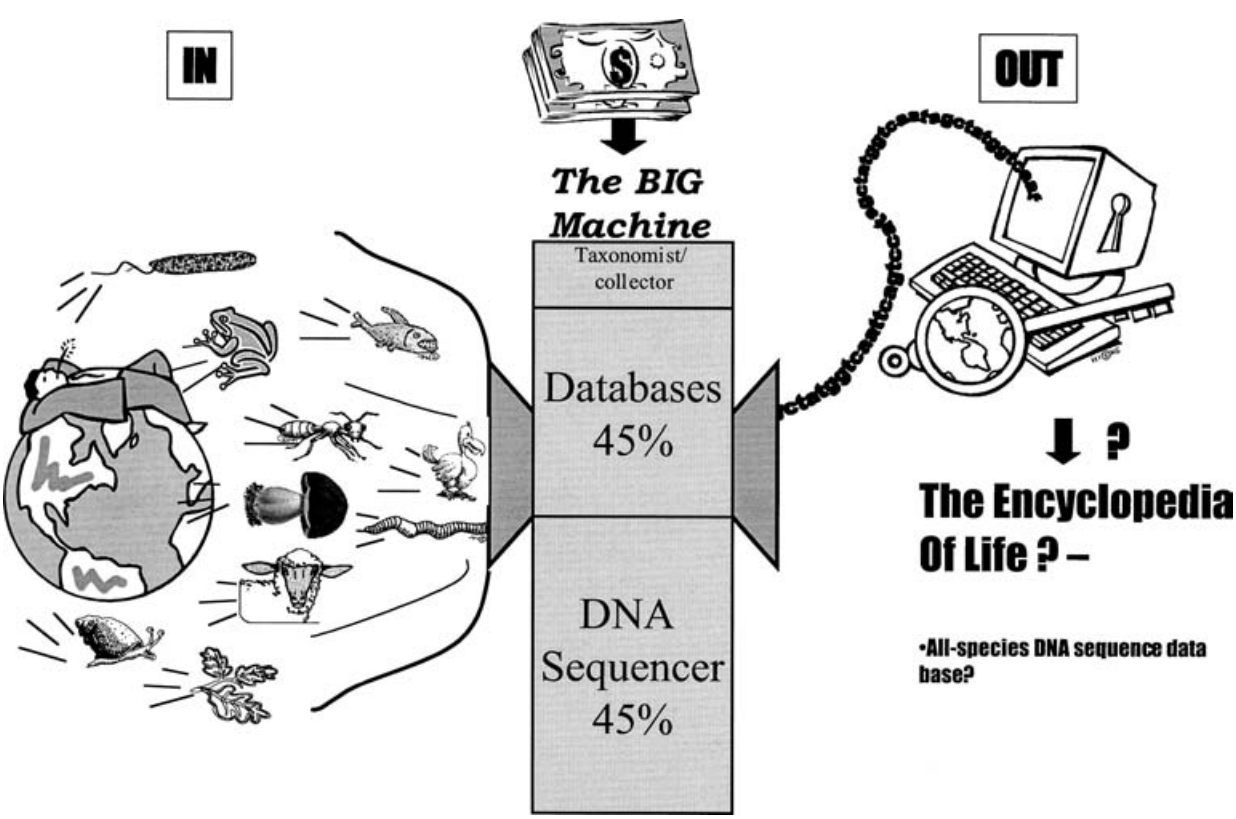

Figure 2 Caricature of the technological solutions of DNA and web-based taxonomy. Images from www.discoveryschool.com and www.barrysclipart.com.

databases facilitating the outputs (Fig. 2), that can automate and speed up the process of assembling the encyclopedia, as well as making taxonomy more objective, and by implication more scientific. One group for which DNA sequence data are indeed productive in discovery, delimitation, description and identification of some sort of taxonomic units is microbial organisms. The vast majority of microbial species have not yet been described, and morphological, physiological or nutritional criteria have failed to converge on a natural (evolutionary) phylogeny (Pace, 1997), or even on useful classifications. It is estimated, for example, that 80–99% of the microorganisms present in soil are unidentified and unculturable (Borneman *et al.*, 1996; Pace, 1997). Aligned DNA sequence data may be used to infer the diversity of mixed microbial communities, where individual organisms are impossible to sample, through phylogenetic trees and multidimensional-scaling plots (Borneman *et al.*, 1996; Olsen *et al.*, 1986), and to discover and define new species through sequence differences (e.g. Rivas *et al.*, 2002). Molecular probes from rRNA have been used *in situ* to identify eukaryotes such as fungi, as well as prokaryotes through a hybridization technique known as oligonucleotide fingerprinting (OFRG; Valinksy *et al.*, 2002*a*, *b*). Bacterial taxonomists using hybridization techniques defined a genospecies on the basis of a DNA–DNA similarity of more than 70% in re-association studies (Amann *et al.*, 1995). More recently, sequencing techniques have shown that organisms that have 70% or greater DNA hybridization similarity will also have at least 96% DNA sequence identity (Stackebrandt & Goebel, 1994).

Sequences from 16S rRNA have now become a standard tool for microbial taxonomy (Hill *et al.*, 2002) and a *Big Machine* for bacterial taxonomy is already well-developed (see diagrams, Amann *et al.*, 1995: 145, 147, 148). The Ribosomal Database Project (Maidak *et al.*, 1999) is designed to facilitate 'species' identification through determining the similarity of a submitted sequence to those already in the database, and large-scale, high-throughput genomic sequencing facilities are also now active using the gene *cpn*60 (Hill *et al.*, 2002). It is expected that developing technologies such as robotic colony picking, template preparation, sequencing and automated data assembly and analysis will be employed to produce comprehensive bacterial community profiles and sequence libraries (Hill *et al.*, 2002). Nevertheless, the amplification of nucleic acids from a community is still likely to exclude individuals and populations. Furthermore, as frequently stressed by bacteriologists, and in stark contrast to the thrust of Tautz *et al.*'s DNA taxonomy proposals, even for very poorly known and characterized groups such as microbes, an rRNA study "*cannot substitute for the isolation and characterisation of microorganisms*" (Amann *et al.*, 1995). Thus, for very poorly known and characterized microbial groups, DNA can indeed provide a rough estimate of some sort of taxonomic units. However *objective* DNA taxonomy may seem, it may not necessarily be useful in terms of identifying species.

None of this implies that DNA taxonomy would be an appropriate or desirable step forward for groups such as plants or animals, which are already comparatively well known, or for taxonomy as a whole. DNA technology has made it possible to close the methodological gap between microbes, and plants and animals (Amann *et al.*, 1995; Olsen *et al.*, 1994). For groups like plants, DNA sequence data can indeed provide an excellent source for placing taxa on a phylogenetic tree. However, there is no evidence to suggest that DNA sequences can be used to discover, delimit, describe and identify species any more quickly than morphological data for most taxa. For many groups the idea that a minute sample of DNA sequence data could replace or even significantly enhance the richly diverse sources of data currently used to discover, delimit and identify species is clearly misplaced. In fact, promoting DNA sequences as the central and essential scaffold for all taxonomy would be an extremely inefficient and retrograde step for most groups.

Sampling

Godfray (2002) contrasts the rapid success of the human genome project with the failure to complete the inventory of species. This is misconstrued. Completing the inventory of biological diversity is a bigger and much more logistically challenging endeavour than sequencing the human genome. Sequencing the genome of a single species is in no way directly comparable to assembling an encyclopedia for 10 million species. A more realistic analogy might be generating DNA sequence data from every human. Even this would be an order of magnitude less complicated than generating sequence data from every species. We do not question that our ability to generate vast quantities of DNA sequence is improving to the extent that large scale *all species* DNA sequencing is a real possibility. However, a proposal to sequence all humans on earth would benefit from a number of factors: (1) we know what constitutes a human; (2) we know where to find them; (3) infrastructure already exists in most countries to sample a significant proportion of people if all governments agreed to do so; and (4) the DNA sequence data so generated would presumably be directly comparable between all humans. In contrast, sampling non-human biodiversity would suffer from three related issues: (1) much of the world's biodiversity (particularly undescribed biodiversity) is difficult to find because of its microscopic size, rarity or elusiveness; (2) there is essentially no infrastructure that exists to sample life on the planet at the scale proposed by Tautz *et al.* (2003); (3) placing DNA sequence data as the central focal point for species discovery, delimitation and identification will be complicated by alignment problems, contamination (by symbionts and parasites, which are themselves of interest but applying the correct sequence data to the correct individual may be very difficult in some cases), horizontal gene transfer, hybridization, lineage sorting between many samples and orthology/paralogy issues.

The issue of sampling, although acknowledged by Tautz *et al.* (2003), has been largely sidelined in the proposals to develop an all-species sequencing project as the basis for DNA taxonomy. Returning to the human genome analogy, sampling many humans on the planet would give us an incredible amount of information about humans but without having the same sort of sampling for other primates we could not identify a human from other species. This is because sequencing one or a few

individuals of a species, or using lots of sequence data from one species but not others, provides little or no information about the underlying variation that may exist within or between species. For example, in a recent paper concerned with species identification of potentially illegally harvested whale meat using the sort of DNA approaches proposed by Tautz *et al.* (2003), Palumbi & Cipriano (1998) discussed the scale of sampling that would be needed and the necessity for highly variable DNA sequence regions. They concluded that even for a small group such as whales, the scale of sampling and technological difficulties precluded a DNA-based approach. For any widespread taxon distributed across numerous regions there is no quick technological fix to achieve a representative sample of variation.

Identification

For any survey in any habitat in any country on the planet, an understanding of biological diversity demands accurate identification in the form of a list of the taxa present. A lack of identification tools of currently described species is perhaps the most immediate bottleneck for field surveys, at least for comparatively well-studied groups (Rejmánek & Brewer, 2001). In this context, a crucial question then is whether DNA sequences can facilitate identification more efficiently than other identification tools. In our view there seem to be very few situations beyond microbes where this is likely to be the case. Accessible identification keys and field guides tailored to a range of users are much more appropriate, important, relevant and urgently required tools than DNA sequence scaffolds to support routine identification tasks. Furthermore, producing workable keys and field guides, although perhaps less glamorous than DNA taxonomy, is an aim that can be completed within the short time frame in which most conservation must be put in place. In addition, making DNA sequences a central requirement for taxonomy will undoubtedly aggravate the already acute North–South inequalities that pervade taxonomic research. North–South collaborations are potentially extremely beneficial in tackling the biodiversity crisis and should be carefully and appropriately nurtured. Finally, facilitating identification *locally* by a wide range of *people* using tools that convey more about a species than simply its name, its DNA sequence and a barcode, are likely to have huge payoffs in stimulating conservation interest and action that will be lost under a system of automated identification carried out by technicians in some distant laboratory.

Conclusions

There seem to be two recent broad-scale justifications for taxonomic research. First, taxonomy is the science underpinning much of conservation (House of Lords, 2002) and biology. Second, that biological diversity is a resource (Wilson, 2000) which like any other asset, can be tapped to enhance human welfare. Ultimately both conservation and rational utilization of the resource demand the completed encyclopedia with all its components (Fig. 1). In essence the real imperative for assembling the encyclopedia of life is an ethical one rather than a purely scientific one. Properly resourced, it can be done, we know how to do it, and it can be made available on the web. Recent funding initiatives in the USA recognize this. However, despite a decade that has seen biodiversity enter the mainstream policy agenda and the establishment of the CBD, not much has changed since May's (1992) paper documenting how little we know about how many species there are. The rate of taxonomic progress is probably about the same (e.g. Prance, 2001) and the time to completion is almost as far away.

The implied and implicit criticism of taxonomists is that we have failed to rise to the challenge of completing the inventory in a realistic period of time, and that this is in part because we persist in old and outdated approaches. There is an element of truth in the image of the taxonomist, often driven by motivations very different from those who prefer to pontificate from the sidelines. However, those of us in the heart of the discipline are often young(ish) and technologically minded, and are actively using database tools and applying DNA sequence data routinely to sort out particular taxonomic groups and complete our bits of the encyclopedia. The collective achievement of taxonomists over the last 200 years is every bit as impressive as the contributions made by related fields of ecology and evolution. Despite this the spotlight remains firmly on what we still do not know about biodiversity, rather than the massive achievements already made. Taxonomy is not in crisis but merely insufficiently resourced to complete the inventory within the time frame that is being imposed by those concerned about the real crisis, which is the biodiversity crisis.

One reason for the perceived lack of progress, is that development of theoretical and technological advances in phylogeny reconstruction, in combination with technological advances in molecular biology, have both driven and consumed much of systematic biology research effort over the last 30 years (Wortley *et al.*, 2002). The overwhelming focus on phylogeny reconstruction among funding agencies, as well as journals that publish high-profile systematic research, has led to tremendous advances in our understanding of the branching order of the tree of life. However, these advances have come at a cost; a lack of interest (i.e. in funding, publication potential and employment possibilities) in the fundamental areas of species identification, description and classification, that are now needed if identification of taxa is indeed a key element in solving the biodiversity crisis. Stark evidence of just how far this agenda has prevailed is provided by the number of gene trees, developed to look at the branching order for particular groups, that have been translated into formal species trees and/or classifications. The answer is relatively few. There are some excellent examples, and certainly quite a few at higher taxonomic levels (e.g. APG, 1998; Tree of Life: www.tolweb.org/tree/phylogeny.html), but, relative to the number of published gene trees, these represent a very small proportion of those being produced for other reasons, or for their own sake. For example, only four out of 20 papers containing gene trees published in two recent volumes of *Systematic Botany* include any, and usually minor, formal nomenclatural changes based on the gene trees. This demonstrates that few gene trees are used in the context of formal nomenclatural changes. It is even more striking to consider how many gene

trees have directly led to the identification of new species. Certainly for plants the examples are few and far between. So why is it that all of these DNA sequence data have not led to vast improvements in both the accuracy and rate of nomenclatural revision and species discovery, as Tautz *et al.* (2003) confidently assume? Clearly there are numerous complications – the need for multiple independent gene trees, organismal and gene sampling issues, lack of resolution and weak support, all mean that gene trees cannot be simply translated into species trees any more than they can instantly delimit or identify species. However, more crucially perhaps, has been the diversion of resources into tree building and away from the inventory. Imposition of DNA taxonomy seems likely to be yet another diversion away from the main task at hand.

The recent history of systematic biology has demonstrated that added technology, as opposed to continued or enhanced investment in taxonomic expertise, does not necessarily result in increased taxonomic productivity. Recent National Science Foundation funding initiatives in the USA recognize this. In the absence of a taxonomic component NSF will no longer fund molecular evolutionary studies under some of its systematics programmes. The two technical solutions, web-based and DNA taxonomy, may provide a tantalizing mirage for politicians concerned about conservation of biodiversity. However, in practice, these ideas are largely a red herring, they do little to address the real problem. The lack of taxonomic progress will not be solved by a requirement for DNA sequences for all groups. In fact it could even detract from the real global priorities for taxonomy, on which many much-maligned taxonomists continue to focus. It is not just that placing the emphasis on DNA sequences will not deliver, but that it will reduce the *encyclopedia* to an impoverished shadow of what the world needs and what E.O. Wilson (Wilson, 2003) is dreaming about. At the end of the day the real reason taxonomists have not yet completed the inventory of biological diversity is that any taxonomic *specialists* worth their salt know that there are no quick answers to the inventory shortfall, and to claim otherwise (Bisby *et al.*, 2002; Godfray, 2002; Lee, 2002; Tautz *et al.*, 2002, 2003) is pie in the sky.

References

Amann, R.I., Ludwig, W. & Schleifer, K.-H. 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiological Reviews* **59**, 143–169.

Angiosperm Phylogeny Group (APG) 1998. An ordinal classification of the families of flowering plants. *Annals of Missouri Botanical Garden* **85**, 531–553.

Bisby, F.A., Shimura, J., Ruggiero, M., Edwards, J. & Haeuser, C. 2002. Taxonomy, at the click of a mouse. *Nature* **418**, 367.

Borneman, J., Skroch, P.W., O'Sullivan, K.M., Palus, J.A., Rumjanek, N.G., Jansen, J.L., Nienhuis, J. & Triplett, E.W. 1996. Molecular microbial diversity of an agricultural soil in Wisconsin. *Applied and Environmental Microbiology* **62**, 1935–1943.

Gewin, V. 2002. Taxonomy: all living things, online. *Nature* **418**, 362–363.

Godfray, H.C.J. 2002. Challenges for taxonomy. *Nature* **417**, 17–19.

Hebert, P.D.N., Cwinska, A., Ball, S.L. & deWaard, J.R. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society Biological Sciences* **270**, 313–321.

Hill, J.E., Seipp, R.P., Betts, M., Hawkins, L., van Kessel, A.G., Crosby, W.L. & Hemmingsen, S.M. 2002. Extensive profiling of a complex microbial community by high-throughput sequencing. *Applied and Environmental Microbiology* **68**, 3055–3066.

House of Lords 2002. *What on Earth? The threat to the science underpinning conservation.* The Stationery Office.

Lee, M.S.Y. 2002. Online database could end taxonomic anarchy. *Nature* **417**, 787–788.

Lipscomb, D., Platnick, N. & Wheeler, Q. 2003. The intellectual content of taxonomy: a comment on DNA taxonomy. *Trends in Ecology and Evolution* **18**, 65–66.

Maidak, B.L., Cole, J.R., Parker, C.T., Jr., Garrity, G.M., Larsen, N., Li, B., Lilburn, T.G., McCaughey, M.J., Olsen, G.J., Overbeek, R., Pramanik, A., Schmidt, T.M., Tiedje, J.M. & Woese, C.R. 1999. A new version of the RDP (Ribosomal Database Project). *Nucleic Acids Research* **27**, 171–173.

Mallet, J. & Willmott, K. 2003. Taxonomy: renaissance or Tower of Babel. *Trends in Ecology and Evolution* **18**, 57–59.

May, R.M. 1992. How many species inhabit the earth? *Scientific American* **267**, 18–24.

Olsen, G.J., Lane, D.J., Giovannoni, S.J. & Pace, N.R. 1986. Microbial ecology and evolution: a ribosomal RNA approach. *Annual Review of Microbiology* **40**, 337–365.

Olsen, G.J., Woese, C.R. & Overbeek, R. 1994. The winds of (evolutionary) change: breathing new life into microbiology. *Journal of Bacteriology* **176**, 1–6.

Pace, N.R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740.

Palumbi, S.R. & Cipriano, F. 1998. Species identification using genetic tools: the value of nuclear and mitochondrial gene sequences in whale conservation. *Journal of Heredity* **89**, 459–464.

Prance, G.T. 2001. Discovering the plant world. *Taxon* **50**, 345–359.

Rejmánek, M. & Brewer, S.W. 2001. Vegetative identification of tropical woody plants: state of the art and annotated bibliography. *Biotropica* **33**, 214–228.

Rivas, R., Velazquez, E., Willems, A., Vizcaino, N., Subba-Rao, N.S., Mateos, P.F., Gillis, M., Dazzo, F.B. & Martinez-Molina, E. 2002. A new species of *Devosia* that forms a unique nitrogen-fixing root-nodule symbiosis with the aquatic legume *Neptunia natans* (L.f.) Druce. *Applied and Environmental Microbiology* **68**, 5217–5222.

Seberg, O., Humphries, C.J., Knapp, S., Stevenson, D.W., Peterson, G., Scharff, N. & Andersen, N.M. 2003. Shortcuts in systematics? A commentary on DNA-based taxonomy. *Trends in Ecology and Evolution* **18**, 63–65.

Stackebrandt, E. & Goebel, B.M. 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic Bacteriology* **44**, 846–849.

Tautz, D., Arctander, P., Minelli, A. & Thomas, R.H. 2002. DNA points the way ahead in taxonomy. *Nature* **418**, 479.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H. & Vogler, A.P. 2003. A plea for DNA taxonomy. *Trends in Ecology and Evolution* **18**, 70–74.

Valinksy, L., Vedova, G.D., Jiang, T. & Borneman, J. 2002*a*. Oligonucleotide fingerprinting of rRNA genes for analysis of fungal community composition. *Applied and Environmental Microbiology* **68**, 5999–6044.

Valinksy, L., Vedova, G.D., Scupham, A.J., Alvey, S., Figueroa, A., Yin, B., Hartin, R.J., Chrobak, M., Crowley, D.E., Jiang, T. & Borneman, J. 2002*b*. Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes. *Applied and Environmental Microbiology* **68**, 3243–3250.

Wilson, E.O. 2000. A global map of biodiversity. *Science* **289**, 2279.

Wilson, E.O. 2003. The encyclopedia of life. *Trends in Ecology and Evolution* **18**, 77–80.

Wortley, A.H., Bennett, J.R. & Scotland, R.W. 2002. Taxonomy and phylogeny reconstruction: two distinct research agendas in systematics. *Edinburgh Journal of Botany* **59**, 335–349.